

# Analyse de données massives (ADM) — **Classification des textes et images**

DUT2 – Semestre 4 – IUT de La Rochelle

# Classification des textes

- **Pré-traitement du texte** : mots en minuscule, suppression des mots vides, lemmatisation, racinisation
- **Représentation de texte** (caractéristiques): TF-IDF, embeddings de mots
- **Division des ensembles de données**: ensembles de train, de test et de validation
- **Algorithmes d'apprentissage automatique**: Naive Bayes, SVM, régression logistique, réseaux de neurones

# Classification des images

- **Pré-traitement d'image** : réduction de taille, couleur RGB, niveaux de saturation HSV, gris
- **Représentation d'image** (caractéristiques): pixels, HOG, SIFT, Canny
- **Division des ensembles de données**: ensembles de train, de test et de validation
- **Algorithmes d'apprentissage automatique**: SVM, régression logistique, réseaux de neurones

# Classification des images

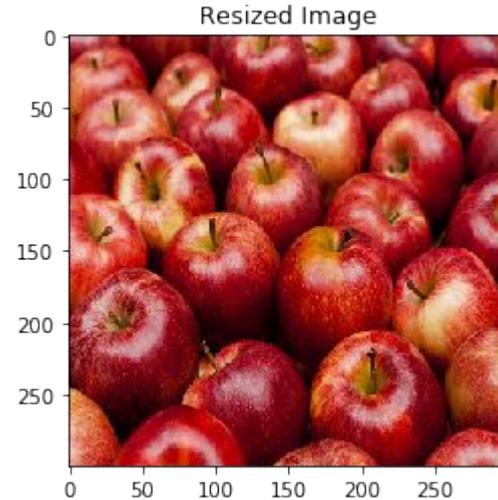
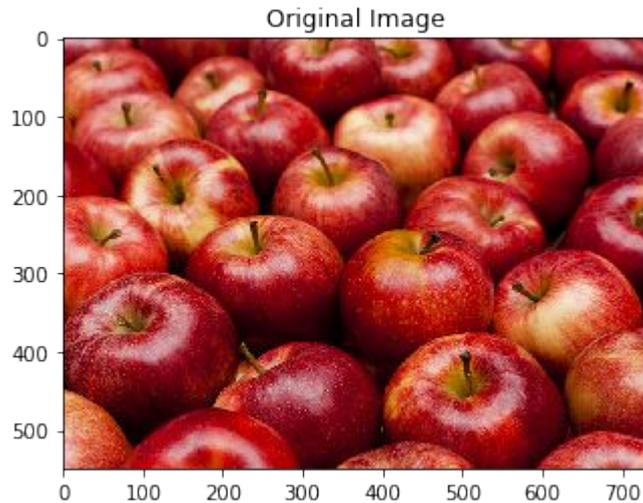
## Pré-traitement

# Pré-traitement des images

- Certains types de représentations fournissent des informations qui ne sont pas évidentes dans une représentation triviale
- Parmi les représentations existantes, nous allons analyser :
  - Gray
  - RGB (Red, Green, Blue )
  - HSV (Hue, Saturation, Value)

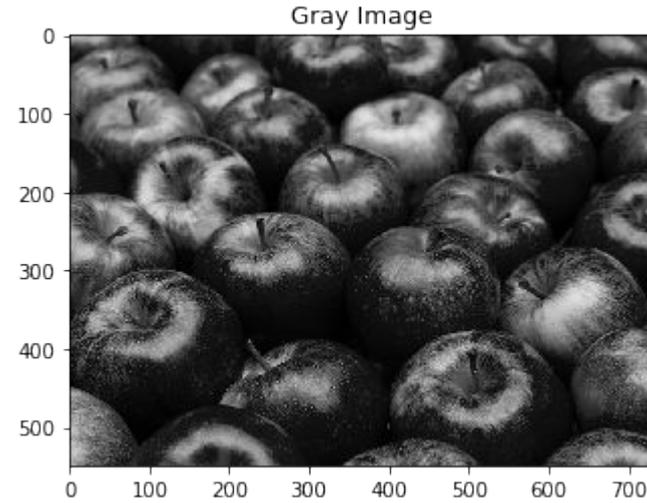
# Pré-traitement des images

- Les images peuvent avoir différentes tailles
- **Redimensionner** les images pour avoir toujours une même taille (par exemple: 100 x 100 pixels)

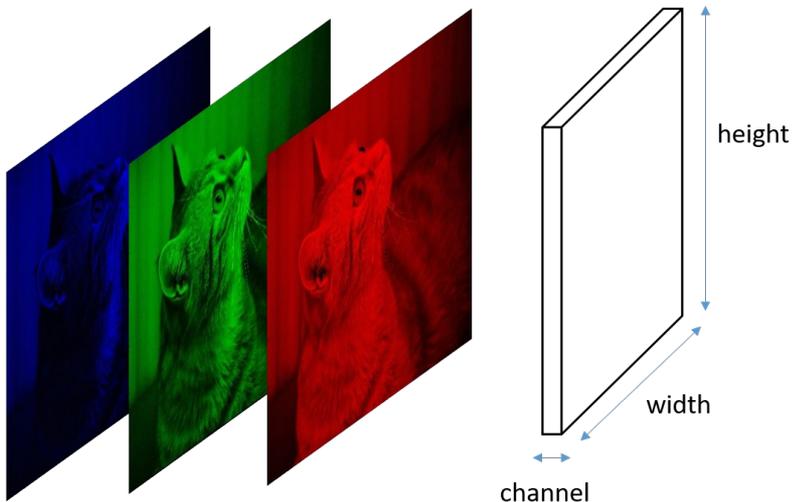


# Pré-traitement des images

- Nous pouvons transformer l'image en **niveaux de gris**



- RGB est un système de codage informatique des couleurs, le plus proche du matériel.
- Le codage RGB indique une valeur pour chacune de ces couleurs primaires.
- 0 - 255 : RVB (rouge, vert, bleu) sont de 8 bits chacun.
- La plage pour chaque couleur individuelle est de 0 à 255 ( $2^8 = 256$  possibilités)



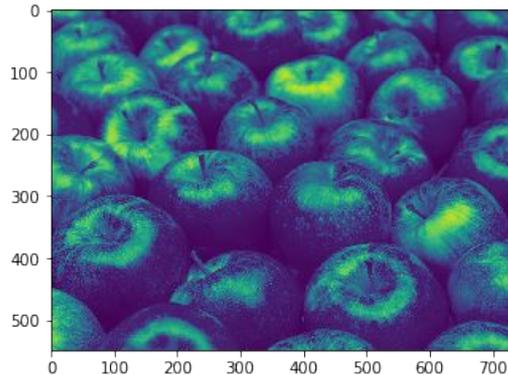
```
In [13]: print("Pixels:", sample_image)
```

```
Pixels: [[[220 169 166]
 [221 170 167]
 [221 170 167]
 ...
 [164 46 58]
 [163 45 57]
 [162 45 54]]

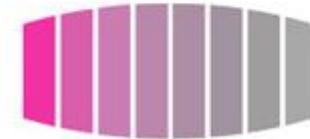
 [[222 171 168]
 [222 171 168]
 [222 173 169]
 ...
 [156 45 54]
 [158 44 54]
 [158 44 54]]

 [[221 172 168]
 [222 173 169]
 [221 173 171]]
```

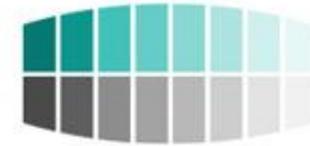
- HSV est une représentation alternative du modèle de couleur RGB. La représentation HSV modélise la façon dont les couleurs apparaissent sous la lumière.
- Teinte (hue) spécifie l'angle de la couleur sur le cercle de couleur RGB.
- La saturation contrôle la quantité de couleur utilisée.
- La valeur (value) contrôle la luminosité de la couleur.



**Hue** is just another another word for color, and it generally refers to a specific slice of the color wheel: the blue slice, the orange slice, the yellow-green slice, and so on.



**Saturation** is the intensity of a color. Highly saturated colors are those that appear bright and pure. Desaturated colors are ones that look dull, muted, or grayed out.



**Value** is the lightness or darkness of a color compared to a scale of grays that goes from near white to near black.

# Classification des images

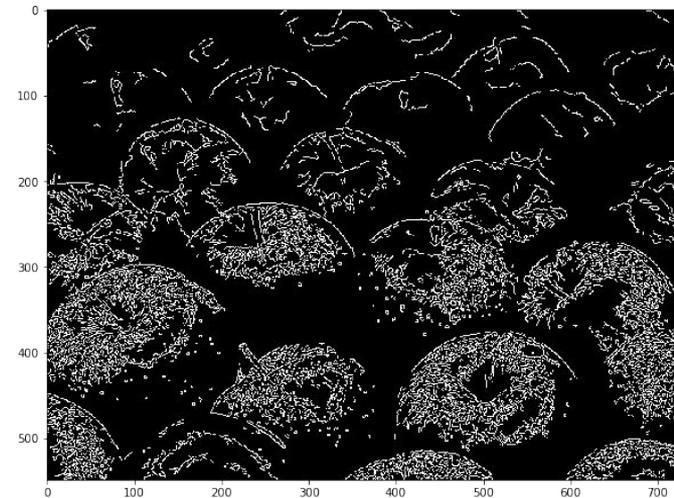
## Représentation des images

# Représentation des images

- Comme dans la première partie où nous avons représenté les mots par des vecteurs, nous allons représenter les images par des vecteurs aussi.
- Comme pour les mots et les documents, il y a plusieurs types de représentations pour les images.
- Chaque représentation peut ajouter des informations supplémentaires sur une image.
- Plus précisément, certains types de représentations fournissent des informations qui ne sont pas évidentes dans une représentation triviale
- Parmi les représentations existantes, nous allons analyser:
  - Canny Edge Detector
  - Gaussian blur
  - HOG (histogram of oriented gradients)
  - SIFT (scale-invariant feature transform)

# Canny Edge Detection (détection de bord Canny)

- est un algorithme de détection de bord populaire. C'est un algorithme pour : la réduction du bruit (étant donné que la détection des contours est sensible au bruit dans l'image, la première étape consiste à supprimer le bruit dans l'image avec un filtre gaussien 5x5.) Et la recherche du gradient d'intensité de l'image.



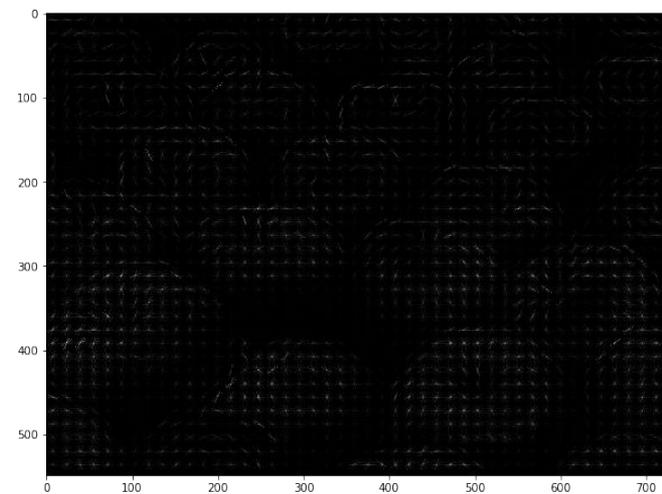
# Gaussian blur : réduire le bruit de l'image

- Dans le traitement d'image, un flou gaussien (également connu sous le nom de lissage gaussien) est le résultat du flou d'une image par une fonction gaussienne.
- C'est un effet largement utilisé généralement pour réduire le bruit de l'image et réduire les détails.



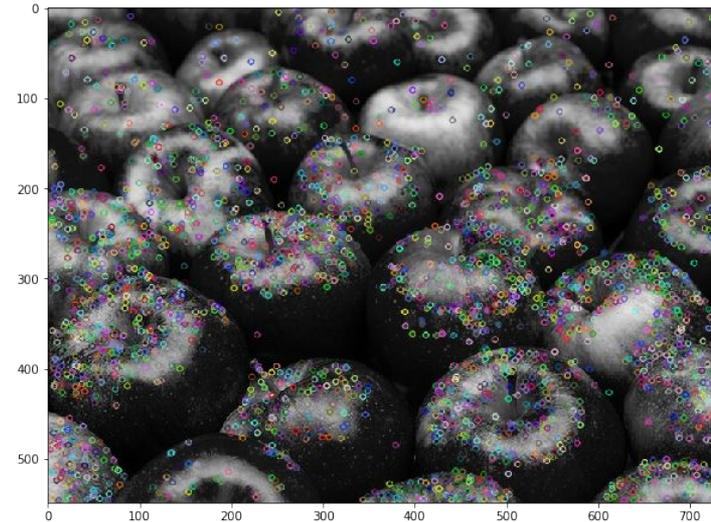
# Histogram of Oriented Gradients (HOG)

- Un histogramme de gradient orienté (HOG) est une caractéristique utilisée en vision par ordinateur pour la détection d'objet.
- La technique calcule des histogrammes locaux de l'orientation du gradient sur une grille dense, c'est-à-dire sur des zones régulièrement réparties sur l'image.



# SIFT (scale-invariant feature transform)

- La transformation de caractéristiques invariantes d'échelle (SIFT) est un algorithme de détection de caractéristiques en vision par ordinateur pour détecter et décrire les caractéristiques locales dans les images.

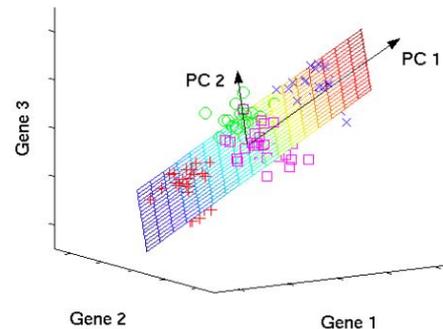
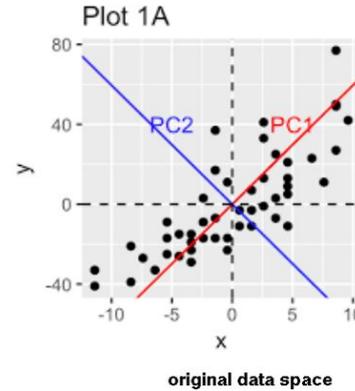


# Représentation des images

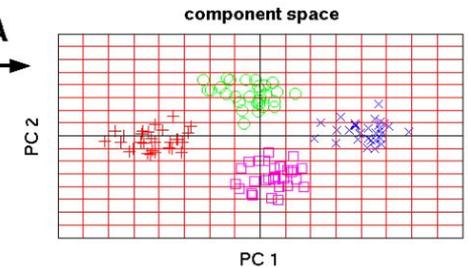
- Nous pouvons utiliser une représentation toute seule ou faire de combinaisons pour ajouter caractéristiques supplémentaires qui sont possibles avec différentes représentations.
- Dans ce cadre, nous pouvons essayer:
  - RGB
  - Gris
  - Gris et SIFT
  - Gris et HOG
  - RGB et Canny
  - et d'autres combinaison
- Pré-traiter les images avec les techniques du traitement d'images (améliorer la luminosité, éliminer le bruit, etc.)

# Principal component analysis (PCA)

- L'analyse en composantes principales consiste à transformer des variables liées entre elles en nouvelles variables décorréélées les unes des autres.
- Ces nouvelles variables sont nommées « composantes principales », ou axes principaux.
- Elle réduit le nombre de variables et rend l'information moins redondante sur une donnée
- Recherche des axes des composantes principales
- Sélection des composantes
- Projection des données sur les axes des composantes



PCA



# Classification des images

## Diviser jeux de données

# Diviser les jeux de données

- Afin d'entraîner les modèles et évaluer la performance de ses modèles avec chaque représentation de mots, nous allons diviser les jeux de données en :
  - entraînement
  - développement (validation)
  - teste



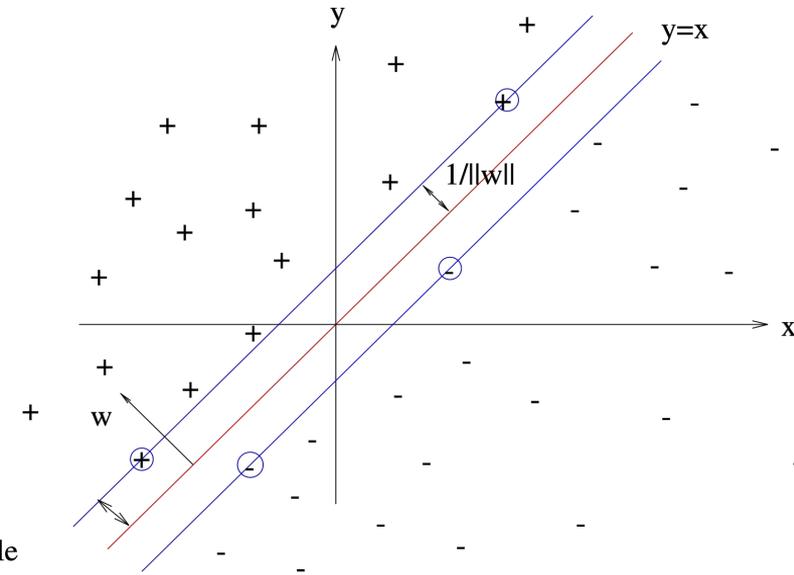
- Afin d'évaluer correctement la performance de chaque modèle, il est très important que les données d'entraînement et de teste soient différentes.

# Classification des images

## Approches de classification

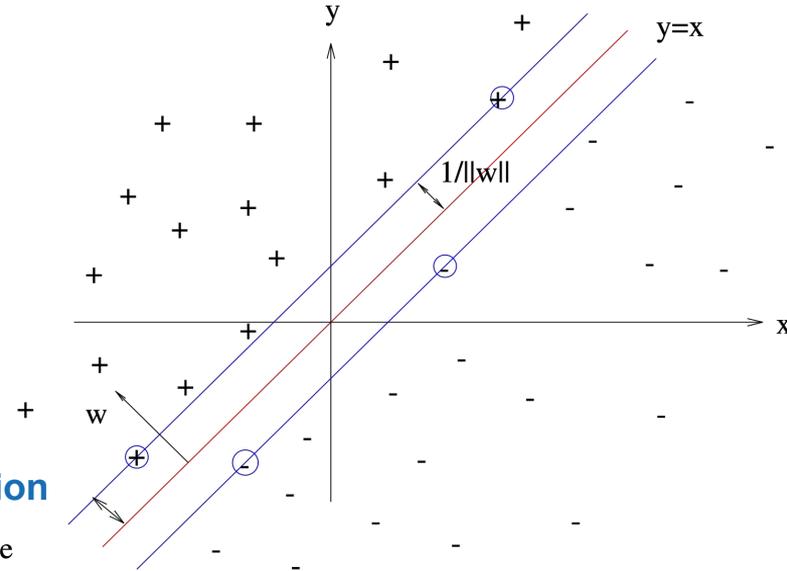
# SVM : support vector machine

- Les SVMs sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination et de régression.
- Ces techniques reposent sur deux idées clés :
  - la notion de marge maximale
  - la notion de fonction noyau
- La marge est la distance entre la frontière de séparation et les échantillons les plus proches.

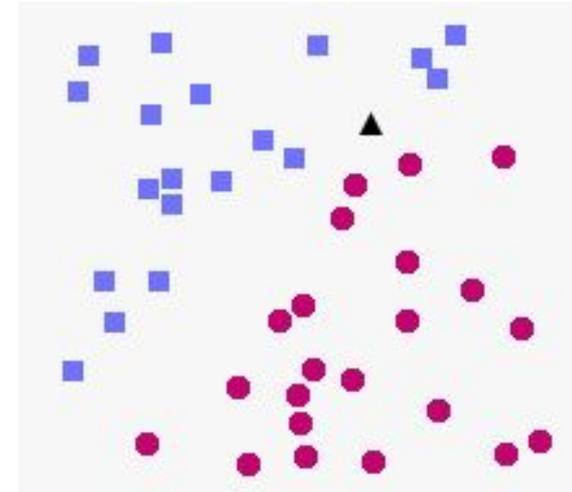


# SVM : support vector machine

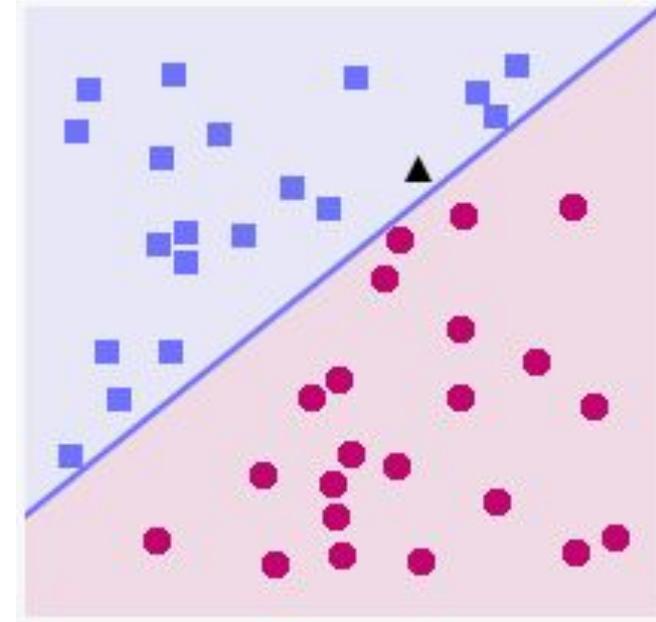
- Les SVMs sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de classification et de régression. Le but est de trouver un «hyperplan» qui pourrait séparer les données avec précision. Il pourrait y avoir de nombreux hyperplans de ce type → «hyperplan optimal» → **problème d'optimisation**
- Ces techniques reposent sur deux idées clés :
  - la notion de «marge» maximale :  
la distance entre la frontière de séparation («hyperplan») et les échantillons les plus proches («vecteurs support»)
  - la notion de fonction noyau («kernel trick») : les données qui ne sont pas linéairement séparables sont transformées pour être mappées dans un nouvel espace  
( $2d \rightarrow \varphi((a, b)) = (a, b, a^2 + b^2) \leftarrow 3d$ , polynomial  $d=2$ )
- Les SVM **maximisent la marge** autour de l'**hyperplan de séparation**
- **Algorithme déterministe !**



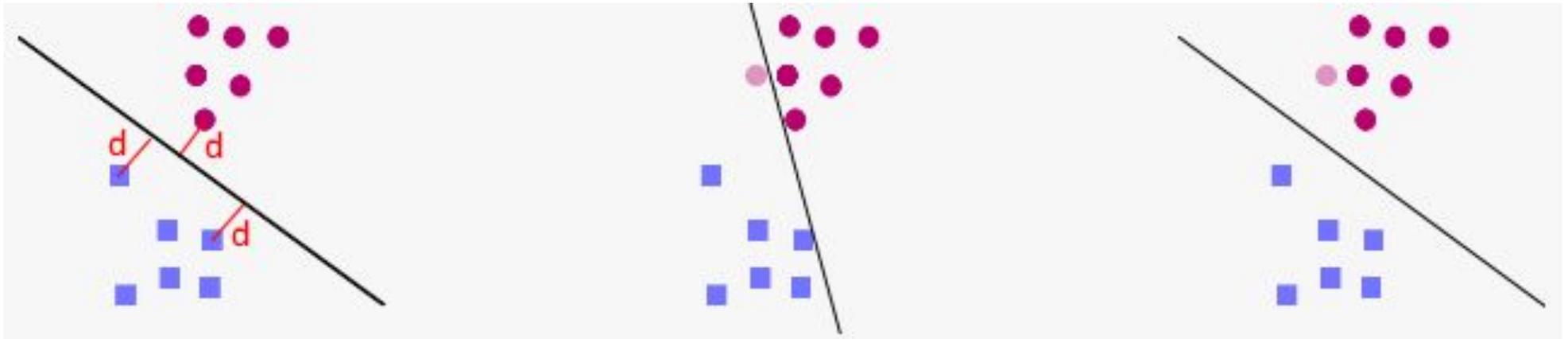
- Problème de classification : pour chaque nouvelle entrée, être capable de déterminer à quelle catégorie cette entrée appartient.
- On se place dans le plan, et l'on dispose de deux catégories :
  - les ronds rouges
  - les carrés bleus
- Cependant, la frontière entre ces deux régions n'est pas connue.
- Le triangle noir est-il un rond rouge ou bien un carré bleu ?



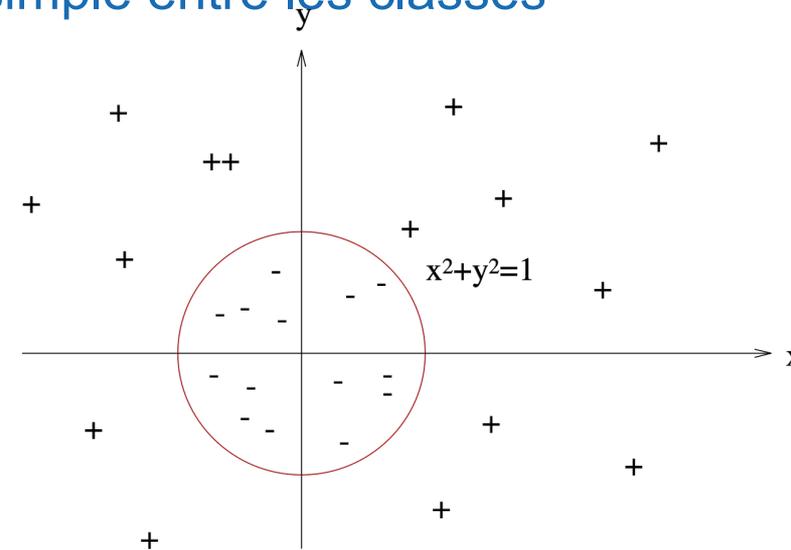
- Avec les jeux de données d'entraînement le SVM va entraîner son modèle à séparer les objets dans deux classes
- Ensuite, il sera capable de prédire à quelle classe appartient un point
- Le triangle noir est en fait un carré bleu.



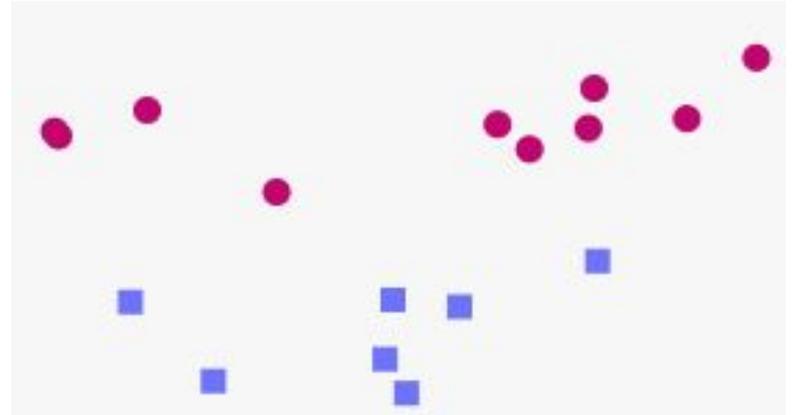
- La marge maximale



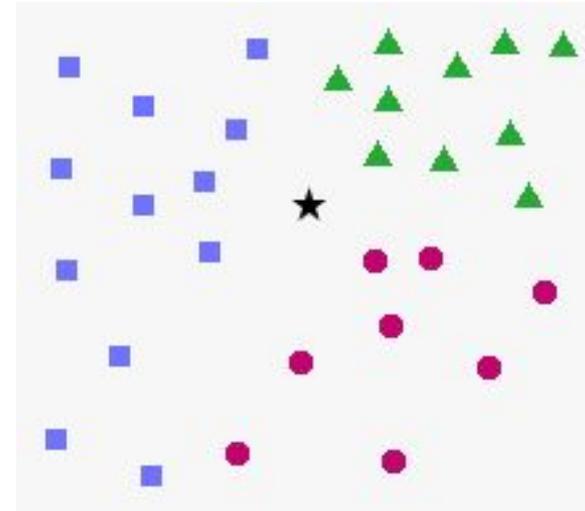
- La deuxième idée est de transformer l'espace de représentation des données d'entrées en un espace de plus grande dimension
- Il est probable qu'il existe une séparation plus simple entre les classes
- Les fonctions noyau permettent de transformer un produit scalaire dans un espace de grande dimension :
  - le noyau polynomial
  - le noyau gaussien



- La fonction noyau

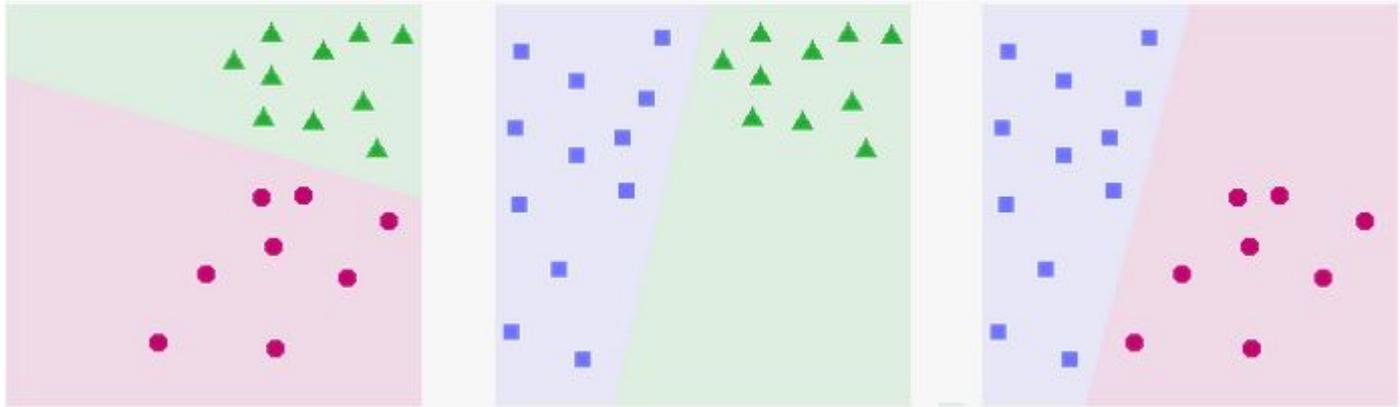


- Classification multi-classes
- SVM : cherche à créer un hyperplan qui sépare l'espace vectoriel en deux:
- Il existe ainsi plusieurs méthodes d'adaptation des SVM aux problèmes multiclassés:
  - One vs one
  - One vs all

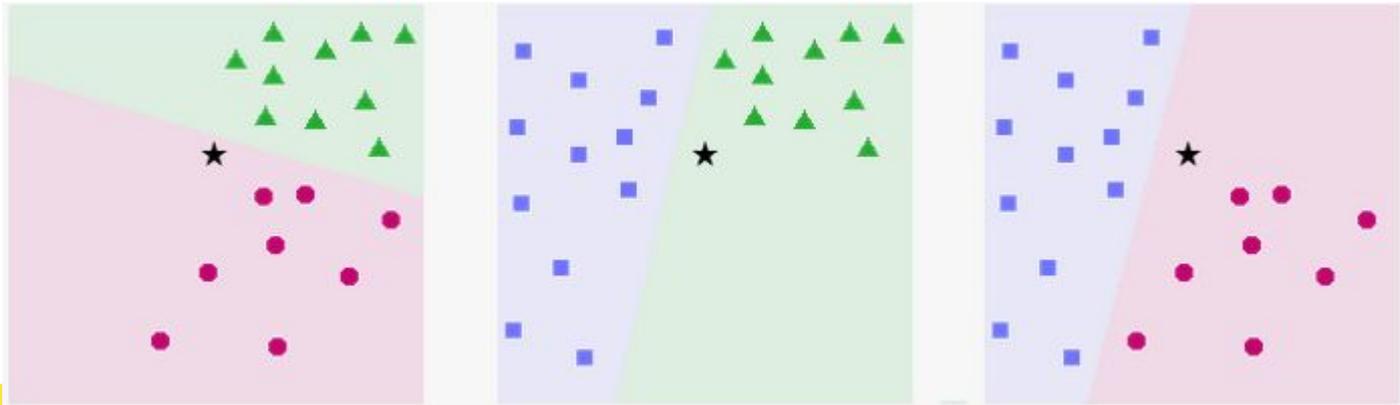


# SVM : One vs one

- Entraînement :

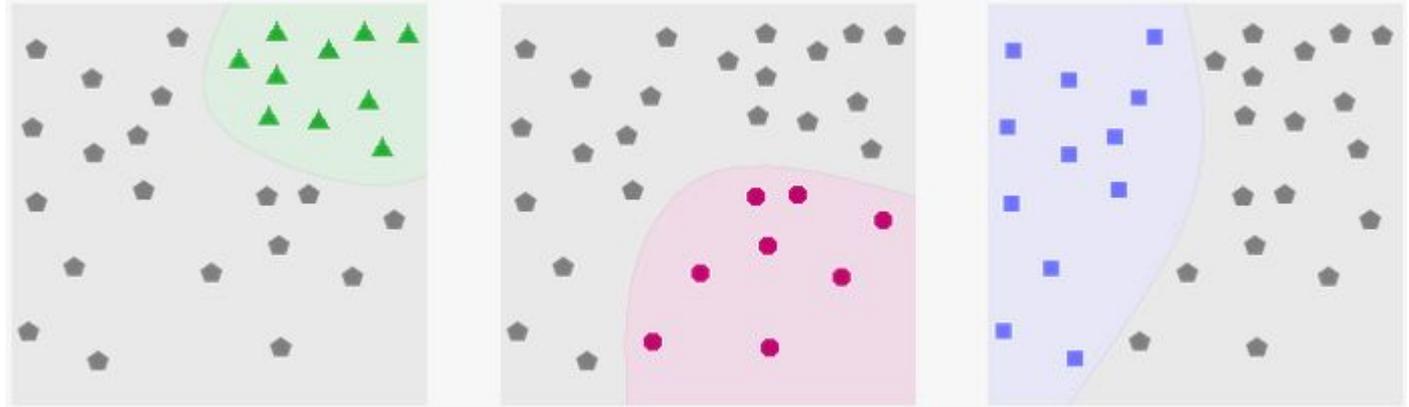


- Prédiction :

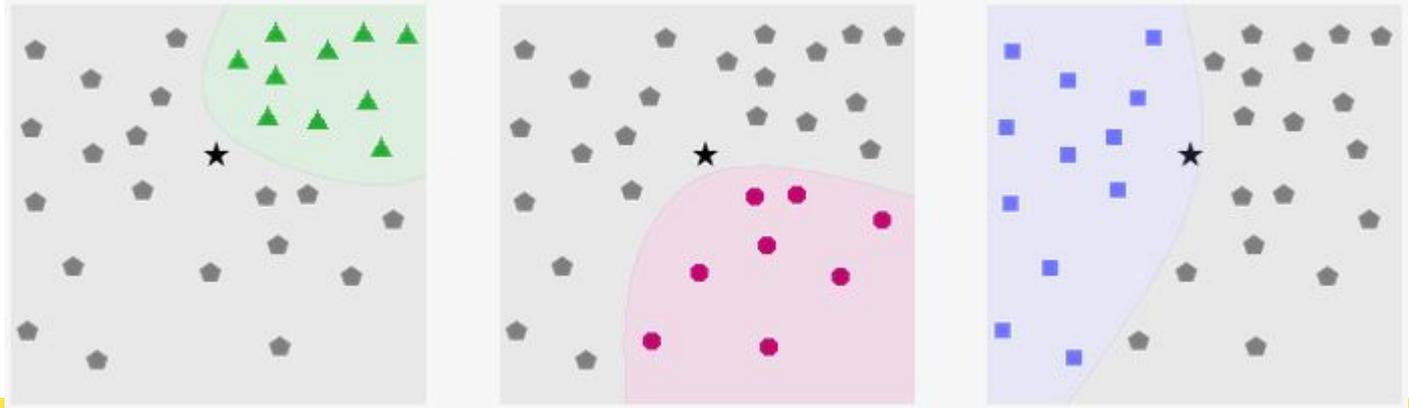


# SVM : One vs all

- Entraînement :



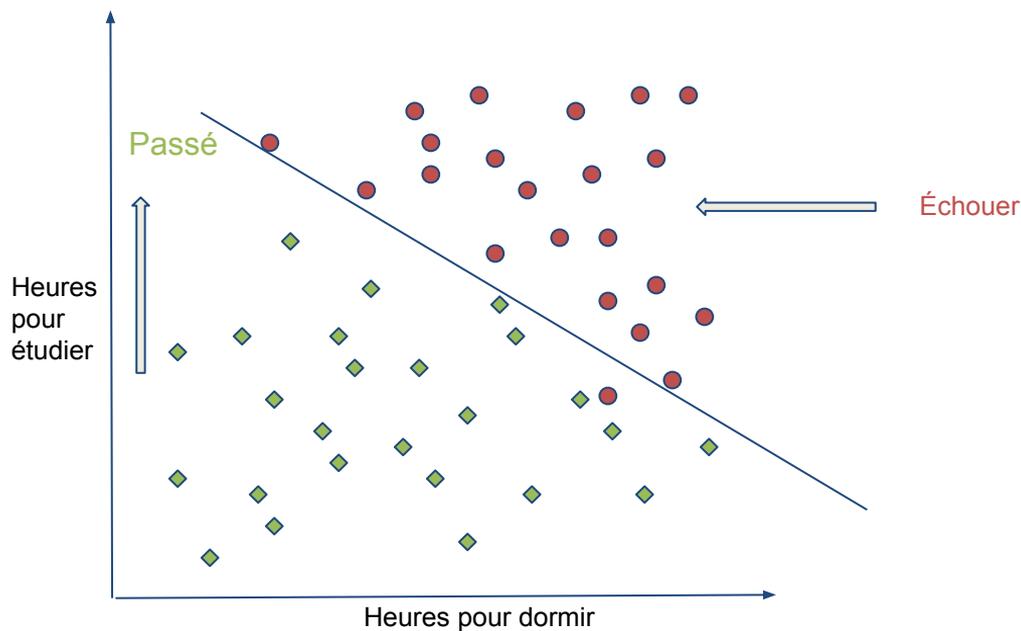
- Prédiction :



# Logistic Regression

- La **régression logistique** est un algorithme de classification qui transforme sa sortie à l'aide de la fonction **sigmoïde** logistique pour donner une valeur de probabilité pour les classes de sortie.
- ~ réseau neuronal à 1 couche
- **Algorithme statistique !**

	Heures pour dormir	Heures pour étudier	Classe
Student 1	8	7	Passé
Student 2	12	5	Échouer
Student 3	10	3	Échouer
Student 4	9	8	Passé
...			



# Classification des images

## Analyse de performance

# Analyse de la performance

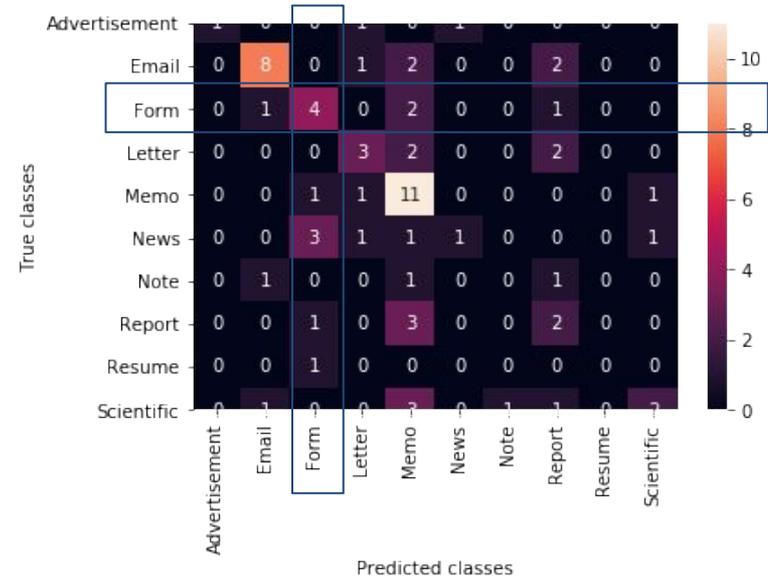
- Après avoir entraîné notre modèle, nous pouvons faire la prédiction des classes des textes du jeux de données de test.
- La fonction f1-score calcule la performance d'une méthode à partir de l'analyse de la quantité de données qui ont été prédit correctement.
  - précision = la proportion des prédictions positifs était effectivement correcte
  - rappel = la proportion de résultats positifs réels a été identifiée correctement
  - F1 = la moyenne pondérée de la précision et du rappel (score final)

# Analyse de la performance

**support** = combien d'images sont Form dans l'ensemble de test (8)  
**recall/rappel** = l'algorithme a détecté 50% des 8 images de type Form  
**precision** = l'algorithme a détecté 10 comme Form (regardez la matrice de confusion). Seulement 40% (précision) sont Form en réalité (4 sur 10).

	precision	recall	f1-score	support
Advertisement	1.00	0.33	0.50	3
Email	0.73	0.62	0.67	13
Form	0.40	0.50	0.44	8
Letter	0.43	0.43	0.43	7
Memo	0.44	0.79	0.56	14
News	0.50	0.14	0.22	7
Note	0.00	0.00	0.00	3
Report	0.22	0.33	0.27	6
Resume	0.00	0.00	0.00	1
Scientific	0.50	0.25	0.33	8
accuracy			0.46	70
macro avg	0.42	0.34	0.34	70
weighted avg	0.48	0.46	0.43	70

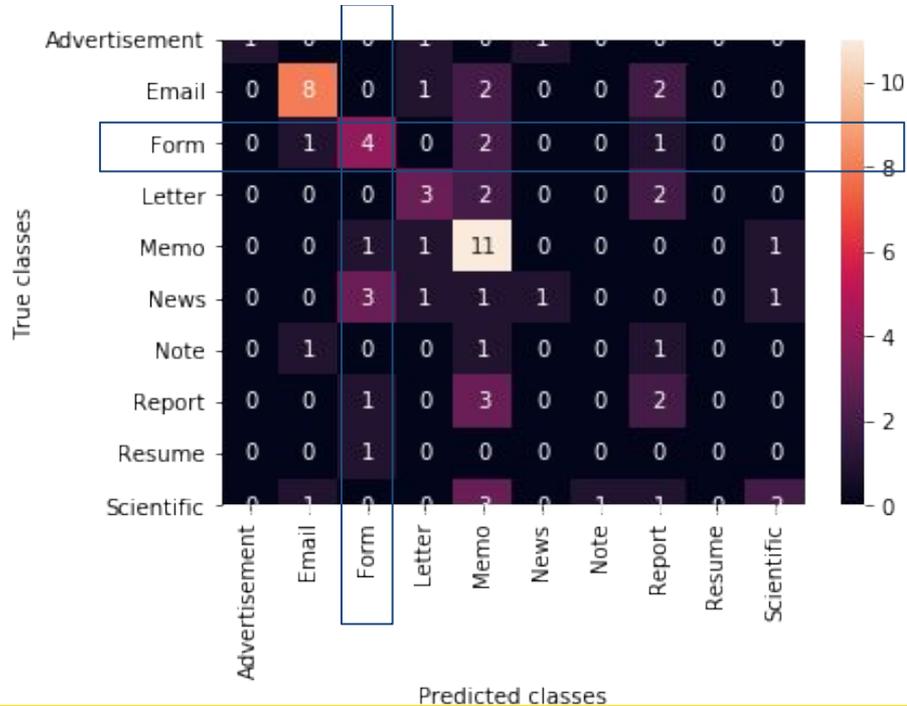
Images Form (8 images):  
 1 a été classé comme Email  
 4 ont été classés comme Form  
 2 ont été classés comme Memo  
 1 a été classé comme Report



# Analyse de la performance

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$



4 vrais positifs = (images Form classées comme Form)

1 (Form classé comme Email) + 2 (Mémo) + 1 (Report) = faux négatifs

1 (Memo classé comme Form) + 3 + 1 + 1 = faux positifs

$$\text{recall}(\text{Form}) = 4/(4+4) = 0.50$$

$$\text{precision}(\text{Form}) = 4/(4+6) = 0.40$$

Form      0.40      0.50      0.44

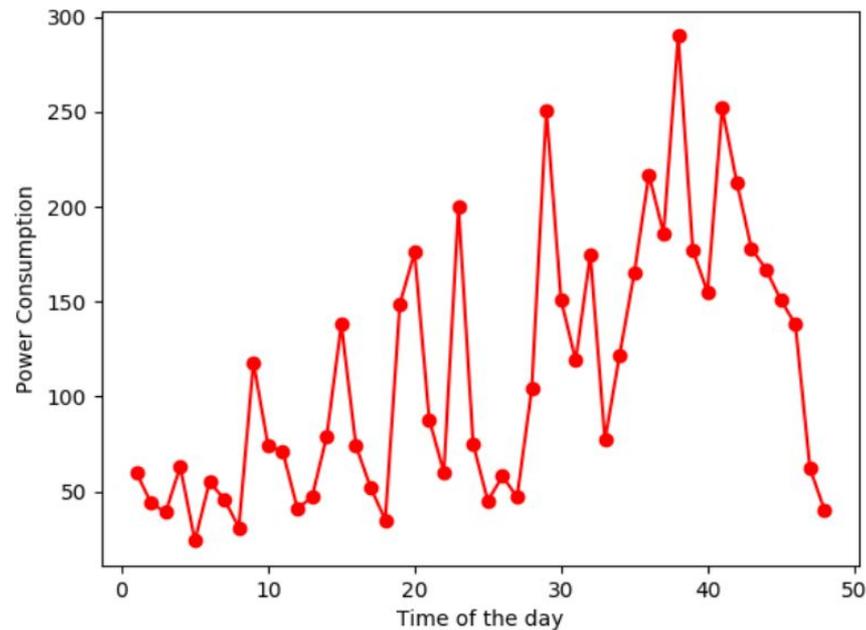
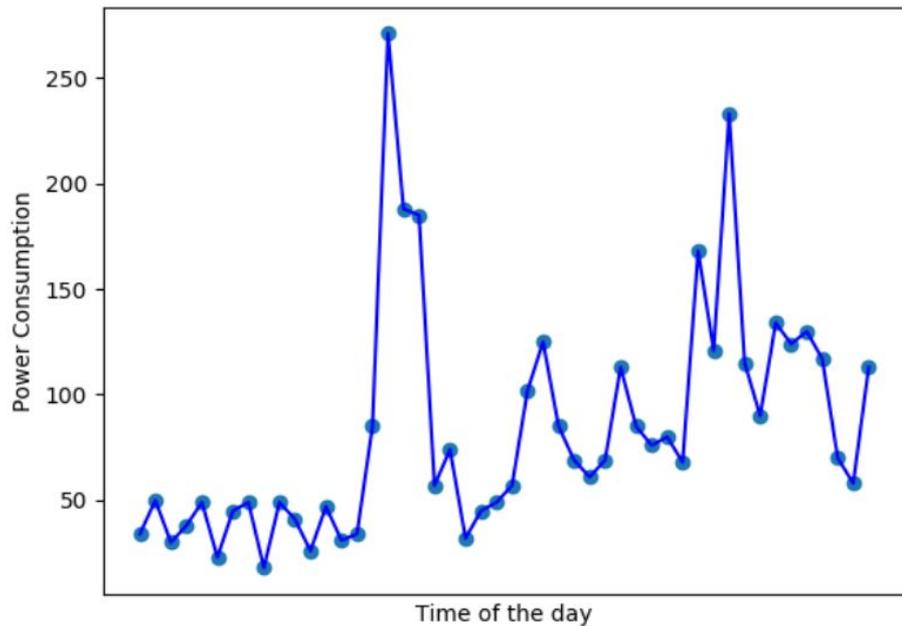
f1(Form) = Moyenne harmonique = 0.44

$$F_1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

# Machine learning

# Machine Learning

- L'apprentissage machine est, après tout, une IA pilotée par les données, et votre modèle ne pourra être qu'aussi bon ou mauvais que les données dont vous disposez.
- En général, vous ne pouvez pas avoir un ensemble de données d'images de voitures et vous attendre à l'utiliser pour classer les chats et les chiens.
- Vous ne pouvez pas utiliser la régression linéaire pour entraîner un modèle sur un ensemble de données qui n'a pas de corrélation linéaire.
- Modèle adapté à la tâche mais si les données ne sont pas bons : le système ne va pas bien marcher



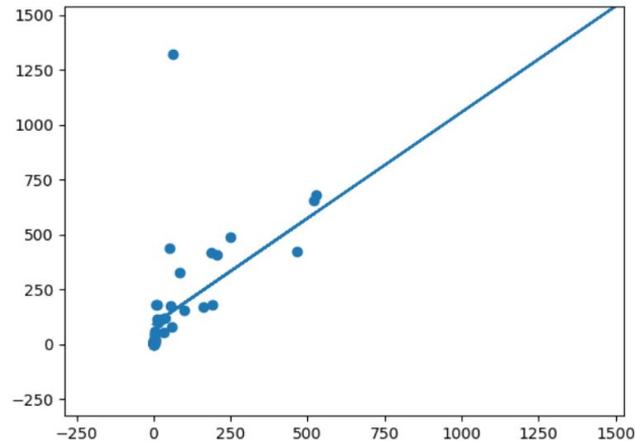
## Machine Learning : Data Correlation

- Vous ne pouvez pas utiliser la régression linéaire pour modéliser un ensemble de données non linéaires.
- L'inverse est également vrai. Si vous avez un ensemble de données corrélées linéaires, vous avez besoin d'un modèle simple comme la régression linéaire.
- Même le meilleur réseau des neurones vous donnera un résultat médiocre.

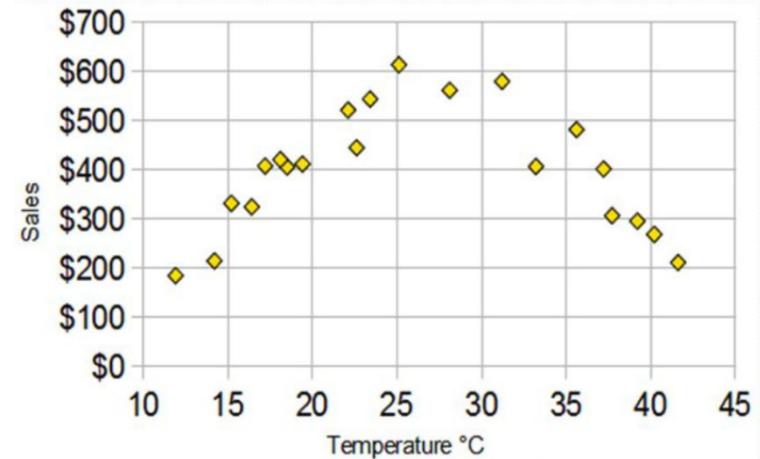
# Machine Learning : Data Correlation

La corrélation des données est la façon dont un ensemble de données peut correspondre à un autre ensemble.

- En ML, pensez à la façon dont vos caractéristiques correspondent à votre sortie.



Brain Weight Vs. Body Weight



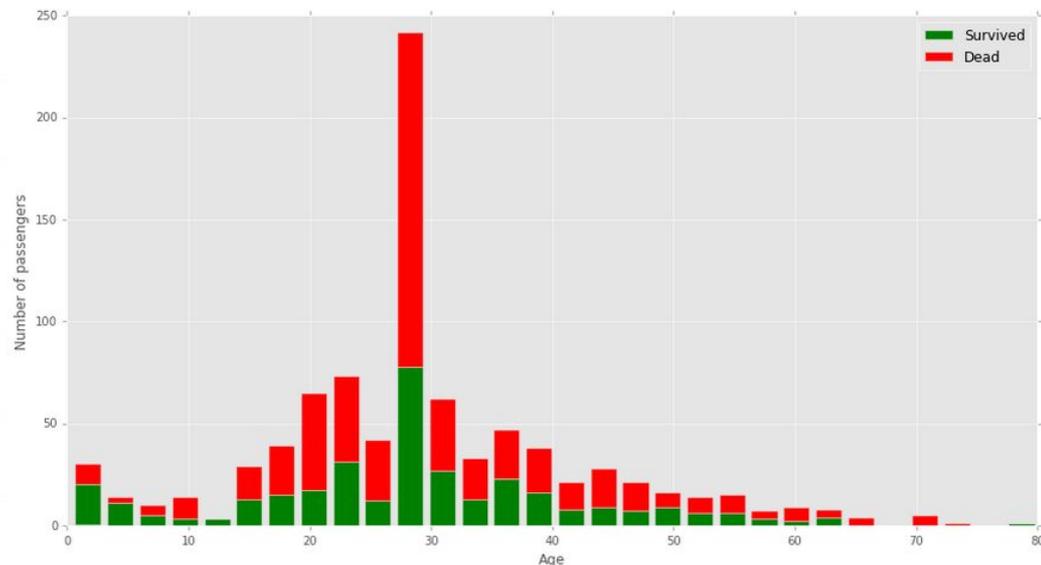
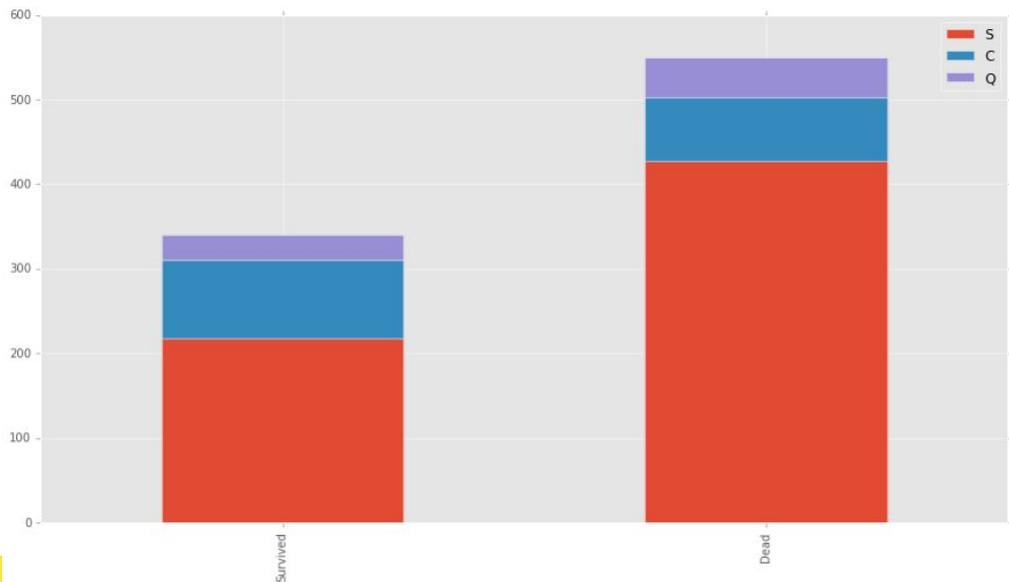
Ice Cream Sales vs. Temperature

## Machine Learning : Filtering data

- Ni tous les données sont pertinentes pour votre projet.
- Dans le cadre du concours Kaggle Titanic, il vous est demandé d'analyser quelles sont les personnes susceptibles de survivre.
- Ce défi comporte de nombreuses caractéristiques.
- Les visualisations peuvent également vous aider à filtrer les caractéristiques inutiles.

# Machine Learning : Filtering data

- Vous pouvez voir la corrélation entre le sexe et la survie, et entre l'âge et la survie.
- A partir de ces visualisations, il devient très évident que le sexe et l'âge ont joué un rôle très important dans la décision de savoir qui aurait pu survivre à bord du Titanic.



# Classification des images Réseaux de neurones

# Réseaux de neurones



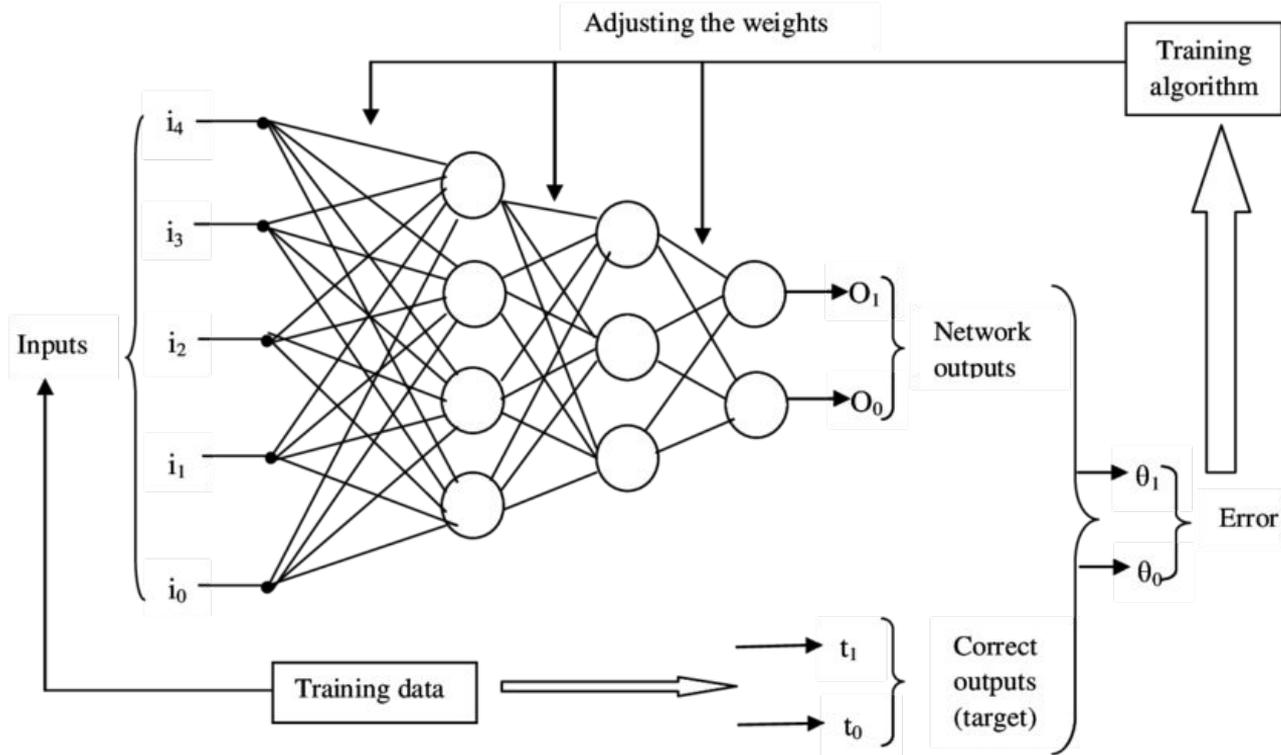
En "x" les entrées, en "y" les sorties. "w" caractérise les paramètres du réseau.



Exemple d'analyse par réseau de neurones.

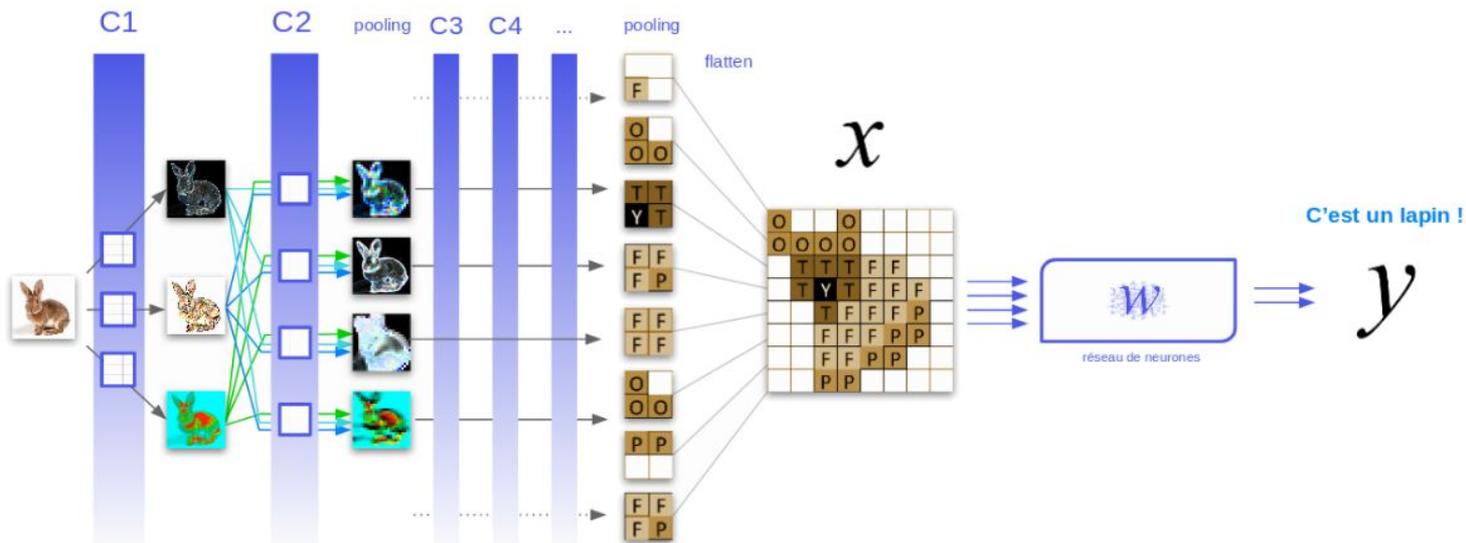
# Réseaux de neurones

(descente de gradient)



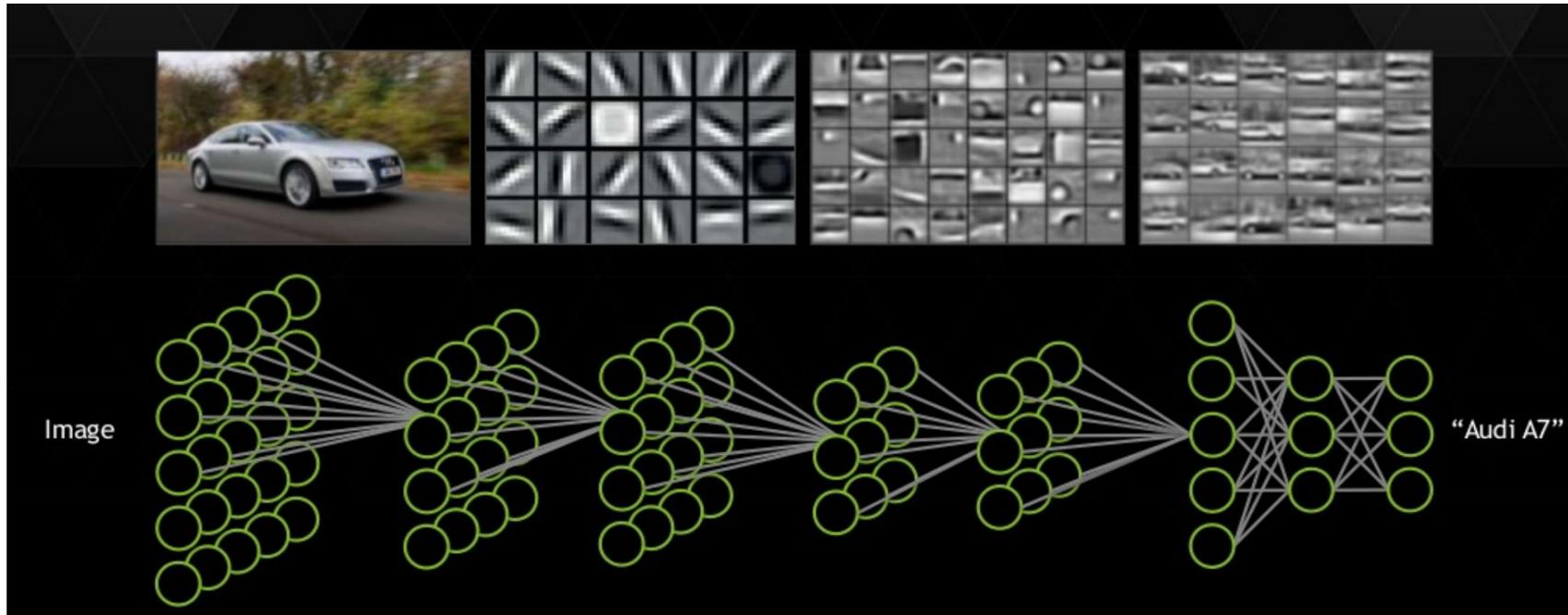
# Réseaux de neurones

- les couches d'un réseau neuronal peuvent détecter automatiquement les contours et autres représentations de vision par ordinateur



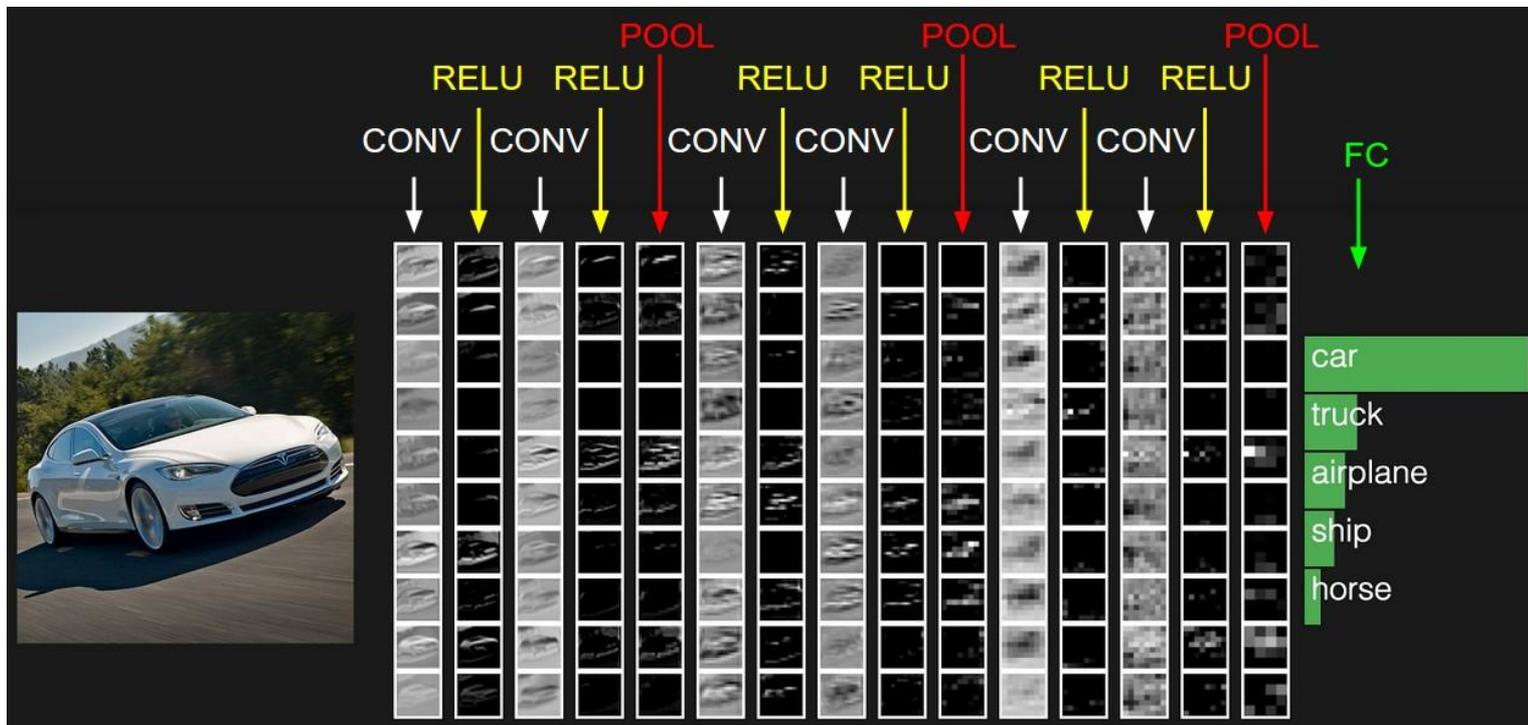
# Réseaux de neurones

- les couches d'un réseau neuronal peuvent détecter automatiquement les contours et autres représentations de vision par ordinateur



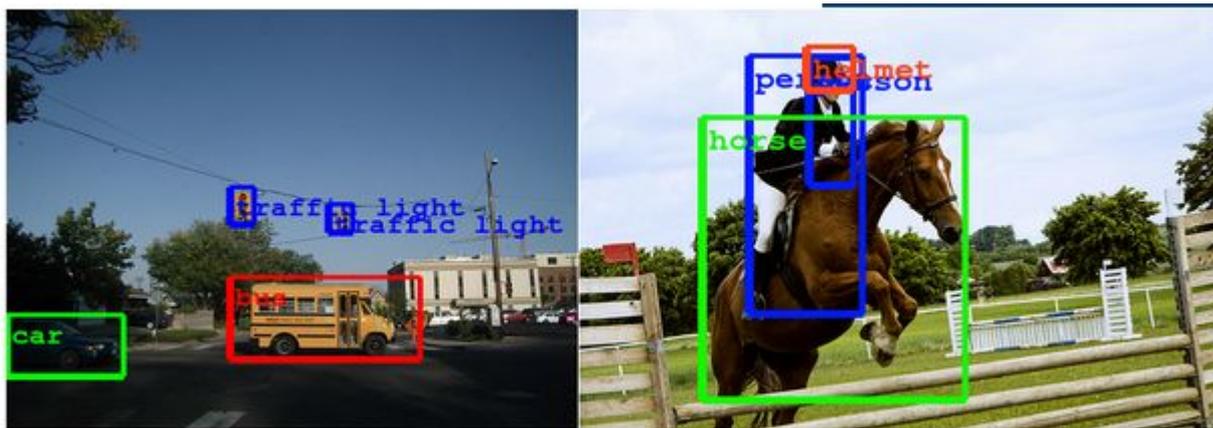
# Réseaux de neurones

- Comment un réseau convolutif voit une image



# Machine Learning : Réseaux de neurones

- <https://youtu.be/r2U-ntB-RM4?t=950>



# Jeux de données : Tobacco3482

Image

Texte

THE TOBACCO INSTITUTE

1875 J STREET, NORTHWEST  
WASHINGTON, DC 20006  
202/457-4800 • 800/898-4433

SAMUEL D. CHILCOTE, JR.  
President

-- VIA FACSIMILE --

September 21, 1994

MEMORANDUM

TO: The Members of the Executive Committee

FROM: Samuel D. Chilcote, Jr. 

Administrative Law Judge Vitonne yesterday presided over the opening day of the Occupational Safety and Health Administration's (OSHA) scheduled 11-week hearings on the proposed indoor air quality standard. OSHA staff testified in the morning; the balance of the day was consumed by a cross-examination period.

Dr. Michael Silverstein, OSHA's director of policy, expressed the Agency's intent to use the information obtained at the hearing to "assure that the final rule is effective." Noting that ETS issues had received a disproportionate amount of pre-hearing attention, Silverstein stressed that tobacco is only one of many airborne "hazards" OSHA would address in the final rule. He denied any political motivation for the rulemaking.

John Martonik, acting director of health standards programs for OSHA, stressed that the regulation would not ban smoking, but only restrict smoking to separately exhausted smoking areas under negative pressure so that nonsmoking coworkers are "protected ... from the adverse health effects of ETS." Martonik estimated that 4.4 million buildings and 70.7 million workers would be affected by the rule. Compliance with the IAQ standard is estimated at \$8.1 billion.

Following the opening statements, nearly a dozen individuals acknowledged their intent to cross-examine OSHA staffers. The questions, primarily from tobacco industry representatives, sought to clarify matters ranging from the scope of the rule to the research the Agency cited to support the proposed action. Following one lengthy exchange, Agency staffers stated that the rule as written does not preempt state and local laws banning/restricting smoking.

CONFIDENTIAL:  
TOBACCO LITIGATION

TICT 0008012

THE TOBACCO INSTITUTE

1875 J STREET, NORTHWEST SAMUEL D. CHILCOTE, JR.  
WASHINGTON, DC 20006 President  
202/457-4800 • 800/898-4433

@- VIA FACSIMILE -@

September 21, 1994

TO: The Members of the Executive Committee

FROM: Samuel D. Chilcote, Jr.

Administrative Law Judge Vitonne yesterday presided over the opening day of the Occupational Safety and Health Administration's (OSHA) scheduled 11-week hearings on the proposed indoor air quality standard. OSHA staff testified in the morning; the balance of the day was consumed by a cross-examination period.

Dr. Michael Silverstein, OSHA's director of policy, expressed the Agency's intent to use the information obtained at the hearing to "assure that the final rule is effective." Noting that ETS issues had received a disproportionate amount of pre-hearing attention, Silverstein stressed that tobacco is only one of many airborne "hazards" OSHA would address in the final rule. He denied any political motivation for the rulemaking.

John Martonik, acting director of health standards programs for OSHA, stressed that the regulation would not ban smoking, but only restrict smoking to separately exhausted smoking areas under negative pressure so that nonsmoking coworkers are "protected ... from the adverse health effects of ETS." Martonik estimated that 4.4 million buildings and 70.7 million workers would be affected by the rule. Compliance with the IAQ standard is estimated at \$8.1 billion.

Following the opening statements, nearly a dozen individuals acknowledged their intent to cross-examine OSHA staffers. The questions, primarily from tobacco industry representatives, sought to clarify matters ranging from the scope of the rule to the research the Agency cited to support the proposed action. Following one lengthy exchange, Agency staffers stated that the

rule as written does not preempt state and local laws banning/restricting smoking.

CONFIDENTIAL:  
TOBACCO LITIGATION

TICT 0008012

Le gouvernement américain a attaqué en justice cinq grands groupes américains du tabac pour avoir amassé d'importants bénéfices en mentant sur les dangers de la cigarette.

Dans ce procès 6 910 192 de documents ont été collectés et numérisés. Afin de faciliter l'exploitation de ces documents par les avocats, vous êtes en charge de mettre en place une classification automatique des types de documents: **Advertisement, Email, Form, Letter, Memo, News, Note, Report, Resume, Scientific.**

# Merci beaucoup pour votre attention!

Si vous avez de questions, vous pouvez nous contacter sur Slack

Aller plus loin:

- <https://openclassrooms.com/fr/courses/4470531-classez-et-segmentez-des-donnees-visuelles>
- <https://blog.octo.com/classification-dimages-les-reseaux-de-neurones-convolutifs-en-toute-simplicité/>
- <https://towardsdatascience.com/data-correlation-can-make-or-break-your-machine-learning-project-82ee11039cc9>
- <https://zestedesavoir.com/tutoriels/1760/un-peu-de-machine-learning-avec-les-svm/>
- <https://blog.betomorrow.com/les-reseaux-de-neurones-de-convolutions-pour-les-ophytes-2b36a59cf648>

# Merci beaucoup pour votre attention!

Si vous avez de questions, vous pouvez nous contacter sur Discord

**Aller plus loin:**

<https://www.kaggle.com/competitions>

Machine learning Coursera famous courses, Andrew Ng,  
<https://www.coursera.org/learn/machine-learning>

Machine learning Coursera (on youtube), Andrew Ng, <https://www.youtube.com/watch?v=PPLop4L2eGk>

The most famous book on deep learning: <https://www.deeplearningbook.org/> (Ian Goodfellow, Yoshua Bengio and Aaron Courville)